DOCUMENT RESUME

ED 431 024 TM 029 857

AUTHOR Segall, Daniel O.
TITLE General Ability Measurement: An Application of Multidimensional Adaptive Testing.
PUB DATE 1999-04-00
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; *Item Response Theory; *Measurement Techniques; Multiple Choice Tests; Scoring; Test Items; Test Use
IDENTIFIERS *Multidimensionality (Tests)

ABSTRACT

Two new methods for improving the measurement precision of a general test factor are proposed and evaluated. One new method provides a multidimensional item response theory estimate obtained from conventional administrations of multiple-choice test items that span general and nuisance dimensions. The other method chooses items adaptively to maximize the precision of the general ability score. Both methods display substantial increases in precision over alternative item selection and scoring procedures. Results of a simulation study based on the Armed Services Vocational Aptitude Battery suggest that the use of these new testing methods may significantly enhance the prediction of learning and performance in instances where standardized tests are currently used. (Contains 5 tables, 1 figure, and 44 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Dan Segall
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 431 024

General Ability Measurement: An Application of Multidimensional Adaptive Testing[1]

Daniel O. Segall
Defense Manpower Data Center
Monterey Bay, CA

Abstract

Two new methods for improving the measurement precision of a general test factor are proposed and evaluated. One new method provides a multidimensional item response theory estimate obtained from conventional administrations of multiple-choice test items that span general and nuisance dimensions. The other method chooses items adaptively to maximize the precision of the general ability score. Both methods display substantial increases in precision over alternative item selection and scoring procedures. Results suggest that the use of these new testing methods may significantly enhance the prediction of learning and performance in instances where standardized tests are currently used.

During most of the 20th century, the role of general cognitive ability (g) for predicting future learning and job-performance has been hotly disputed. Schmidt and Hunter (1998) summarize 85 years of validity research, stating that the most well known conclusion from this work is that "for hiring employees without previous experience in the job the most valid predictor of future performance and learning is general mental ability" (p. 262). Ree and Earles (1991a, 1992, 1994) arrive at a similar conclusion for the prediction of both training and job success. These conclusions emphasize the importance and usefulness of well constructed measures of general ability. It follows that more precise measures of g will lead to a number of desirable personnel selection outcomes, including increases in employee performance, and increased learning of job-related skills (Hunter, Schmidt, & Judiesch, 1990).

The precursor to modern ability theory (Spearman, 1904) states that the variation in error-free mental measurements is due to two factors. One factor, termed general cognitive ability is common to all mental ability measurements, while the other factors s are test specific. Given that s and g are uncorrelated, and the s's are uncorrelated with each other, Spearman's two-factor formulation suggests that a composite score formed from a number

[1] This paper was presented at the annual meeting of the National Council on Measurement in Education (April, 1999), Montreal Canada. The views expressed are those of the author and not necessarily those of the Department of Defense, or the United States government.

Requests for copies should be sent to: Daniel O. Segall, Defense Manpower Data Center, DoD Center Monterey Bay, 400 Gigling Road, Seaside, CA 93955-6771. Email: segalldo@osd.pentagon.mil

TM029857

BEST COPY AVAILABLE

of tests will have more *g* than any of the individual components used to form the composite. Although Spearman's two-factor theory has been largely supplanted by hierarchical models of intelligence (which allow for the existence of intermediate group factors), *g* is widely assumed by modern cognitive theorists to exist at the apex of all mental measurements. Accordingly the prevailing approach to general cognitive ability measurement is consistent with Spearman's original formulation. This approach attempts to average out much of the variance due to the unique demands made by each test or test-grouping by computing composite scores from a large number of highly diverse tests.

Since Spearman's early insight, progress towards the construction of a perfect measure of general cognitive ability has been disappointing. The amount of *g* variance contained in the best standardized measures currently in use may be as low as 64 to 75 percent of the total variance (Jensen, 1998, p. 309; Ree & Earles, 1991b). This lack of progress may seem somewhat contradictory in light of substantial test-theory advances made during this period (see Cronbach, 1970; Gulliksen, 1987; Horst, 1966; Lord, 1980; Lord & Novick, 1968; Thurstone, 1947; van der Linden & Hambleton, 1997). Using modern test construction techniques, very precise measures of narrowly defined abilities, proficiencies, or knowledge domains can be constructed—measures which correlate .95 or greater with the true underlying proficiency. For example, item response theory (IRT) techniques can be used to construct highly precise scales of unidimensional domains. Unfortunately these scales and associated domains tend to contain high levels of specificity, and the problem originally addressed by Spearman remains: How can specific variance contained in individual tests be removed to obtain a precise measure of general ability? Thus the precise measurement of narrowly defined abilities as afforded by modern test theory does not guarantee precise measurement of the general underlying cognitive ability.

To deal effectively with the contributions of systematic error, Humphreys (1981, 1985, 1986) has advocated measures containing a variety of unwanted or specific variance. However, practical concerns over test-length and efficiency have placed limits on the numbers and diversity of tests contained in typical selection batteries. For example, general ability scores are often computed from a composite consisting of a small number of tests (e.g. math, verbal, spatial, and reasoning). Consequently, the specific errors associated with unique test demands tend not to cancel to a sufficient degree.

What is needed then is an approach to cognitive ability estimation that can systematically remove specific-factor variance from the general ability estimate. One set of approaches are based on factor score estimation procedures—a collection of methods for expressing factors in terms of the observed test scores (Harman, 1976, Chapter 16). These methods form an estimate of *g* from linear combinations of test scores. In practice these methods tend to produce factor score estimates that are highly correlated with each other, and with unit weighted composites (Ree & Earles, 1991b). These findings are consistent with Wilks' (1938) theorem which states that correlations among differentially weighted composites tend toward one as the number of positively correlated tests in the composite increases. These factor score estimation procedures are rarely used in practice since they tend to produce only marginal or trivial gains in precision over less complex unit weighting procedures.

An appealing alternative approach for the estimation of general-ability models *item* (rather than *test*) variables in terms of uncorrelated general and specific (or nuisance)

dimensions. The nuisance dimensions carry irrelevant variance which inflates and distorts the item-variance. By partialling out this unwanted variance at the item, rather than test-score level, the precision of ability estimation can be greatly enhanced, depending on the level of influence of the nuisance dimensions on the item-response variables. As with existing methods, accurate estimation of general ability requires a sufficiently broad representation of nuisance dimensions by test-items in the battery. However, since the effects of nuisance dimensions on item-responses are modeled explicitly, accurate measurement of g by this approach does not require random cancelation of specific errors across a large number of test scores.

Two new methods of general ability measurement are presented which provide a mechanism for removing unwanted item variance. There are several fundamental characteristics of the approaches presented here which distinguish them from traditional methods of general cognitive ability measurement. First, unlike traditional approaches that are based on classical test theory, both proposed approaches are based on multidimensional item response theory techniques. Second, both approaches assume a residualized form of the underlying factor model. That is, these methods partial out the effects of unwanted variance from the item response variables. This residualized form of the model expresses each item variable in terms of uncorrelated general and nuisance dimensions. A third distinguishing feature of one proposed approach is that items are selected adaptively to maximize the precision of the general ability score.

Although compelling arguments for the superiority of the proposed methods can be made on theoretical grounds, the level of improved measurement efficiency may not outweigh their additional computational complexity. To estimate likely benefits achieved by a realistic application, levels of precision for proposed and alternative techniques are evaluated from simulated response data based on a high-stakes high-volume personnel selection battery.

Item Response Model and Ability Estimation

Many constructs measured by psychological tests can be conceptualized in terms of a hierarchy of abilities. Table 1 displays a set of equations for a hypothetical three-level factor model, where each ability η_j is expressed in terms of two sources θ_j and η_k (for $j = 1, ..., p$; $k = 1, ..., p$; $j \neq k$). The θ_j are random ability variables which are uncorrelated with each other. The η_j are functions of θ_j and other η_k. At the highest level is general ability denoted by $\eta_1 (= \theta_1)$. At the next highest level are verbal and math abilities (η_2 and η_3) which are assumed to be linear combinations of general η_1 and specific (θ_2, θ_3) abilities. At the lowest level are factors for AR, WK, PC, and MK abilities which are assumed to be linear combinations of verbal or math abilities (η_2 and η_3) and specific factors $\theta_4, \theta_5, \theta_6, \theta_7$.

The general hierarchical model can be succinctly represented in matrix notation by

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\theta} \,, \tag{1}$$

where $\boldsymbol{\eta}$ is a $p \times 1$ vector of ability variables, $\boldsymbol{\theta}$ is a $p \times 1$ vector of specific random ability variables, and $\mathbf{B}$ is the $p \times p$ coefficient matrix describing the influence of the latent ability variables on each other. Here, $\mathbf{B}$ is a lower-triangular matrix with the main diagonal equal to zero, $\mathrm{E}(\boldsymbol{\theta}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\theta}\boldsymbol{\theta}') = \mathbf{I}$, and $\mathbf{I}$ is a $p \times p$ diagonal matrix with all diagonal elements

Table 1: Three level hierarchical factor model.

Level	Ability	Equation
3	General	$\eta_1 = \theta_1$
2	Verbal	$\eta_2 = b_{21}\eta_1 + \theta_2$
2	Math	$\eta_3 = b_{31}\eta_1 + \theta_3$
1	AR: Arithmetic Reasoning	$\eta_4 = b_{43}\eta_3 + \theta_4$
1	WK: Word Knowledge	$\eta_5 = b_{52}\eta_2 + \theta_5$
1	PC: Paragraph Comprehension	$\eta_6 = b_{62}\eta_2 + \theta_6$
1	MK: Math Knowledge	$\eta_7 = b_{73}\eta_3 + \theta_7$

equal to unity. Solving (1) algebraically for $\boldsymbol{\eta}$, we arrive at the reduced form

$$\boldsymbol{\eta} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\theta}\,. \tag{2}$$

The covariance matrix among latent factors is given by $\Phi \equiv \mathrm{E}(\boldsymbol{\eta}\boldsymbol{\eta}') = \mathbf{T}\mathbf{T}'$, where $\mathbf{T} = (\mathbf{I} - \mathbf{B})^{-1}$.

In modeling responses to a battery of n individual items, we assume an underlying continuous variable z_i for each item i $(i = 1, ..., n)$. Continuing with our example, let us assume that there are a total of 105 items—each item variable loading on one of the four first-order factors $(\eta_4, \eta_5, \eta_6, \eta_7)$ through the relations:

$$\text{AR}\begin{cases} z_1 = \gamma_{1,4}\eta_4 + \epsilon_1 \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ z_{30} = \gamma_{30,4}\eta_4 + \epsilon_{30} \end{cases} \tag{3}$$

$$\text{WK}\begin{cases} z_{31} = \gamma_{31,5}\eta_5 + \epsilon_{31} \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ z_{65} = \gamma_{65,5}\eta_5 + \epsilon_{65} \end{cases} \tag{4}$$

$$\text{PC}\begin{cases} z_{66} = \gamma_{66,6}\eta_6 + \epsilon_{66} \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ z_{80} = \gamma_{80,6}\eta_6 + \epsilon_{80} \end{cases} \tag{5}$$

$$\text{MK}\begin{cases} z_{81} = \gamma_{81,7}\eta_7 + \epsilon_{81} \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ z_{105} = \gamma_{105,7}\eta_7 + \epsilon_{105} \end{cases}. \tag{6}$$

Note that each item variable is assumed to be a function of a single first-order factor and a random error term ϵ_i. The general relation can be summarized in matrix notation by

$$\mathbf{z} = \Gamma\boldsymbol{\eta} + \boldsymbol{\varepsilon}\,, \tag{7}$$

where $\mathbf{z} = \{z_1, ..., z_n\}'$, Γ is a $n \times p$ coefficient matrix, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of random error variables which are uncorrelated with each other, and with $\boldsymbol{\eta}$.

When partitioning the latent factors into general and secondary (or nuisance) dimensions, a useful alternative parameterization for the item variables $\mathbf{z}$ is obtained by substituting (2) into (7) which provides

$$\begin{aligned} \mathbf{z} &= \Gamma(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ &= \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,, \end{aligned} \tag{8}$$

where

$$\mathbf{A} = \Gamma(\mathbf{I} - \mathbf{B})^{-1}\,. \tag{9}$$

In (8), the latent item variables $\mathbf{z}$ are expressed directly in terms of uncorrelated abilities $\boldsymbol{\theta}$. As described in the following section, this parameterization allows the focus of measurement to be shifted away from the first-order correlated dimensions $(\eta_4, \eta_5, \eta_6, \eta_7)$ and directly to the general θ_1-dimension.

Given the hierarchical features of the model, a particular pattern of zero and nonzero parameters emerges for the **A**-matrix. This pattern can be examined by setting appropriate elements of the **B** and Γ matrices equal to 1 and by performing the matrix calculations displayed on the right-hand side of (9). The pattern of free and fixed (zero) parameters for the example defined by Eqs. (3) through (6) and Table 1 is displayed in Table 2. This pattern resembles an extension of Holzinger's (1937) bi-factor model, where each item-variable loads on three (rather than two) dimensions. McDonald (1985, pp. 105–107) suggests a similar pattern of loadings for conducting hierarchical factor analyses. In general, if **B** and Γ are known, then any hierarchical model of the form given by (2) and (7) can be reparameterized by (8)[2].

Given a set of tenable assumptions (Lord, 1980, p. 31), an item response model can be formulated to describe the relation between the continuous item-variables $\mathbf{z}$ and the dichotomous item responses $\mathbf{u} = \{u_1, u_2, ..., u_n\}'$. According to this model, if the value of z_i is larger than some item-specific threshold δ_i, then the item is answered correctly, otherwise it is answered incorrectly. This model leads to the normal ogive response model (Lord, 1952), or alternatively to an approximation given by the logistic function (Birnbaum, 1968). The effects of guessing can be incorporated into the model through the use of a lower asymptote.

The multidimensional logistic form of this item-response model (Hattie, 1981) postulates p latent traits $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_p\}'$, where each trait affects performance on one or more items. The observed data consist of a vector of scored responses $\mathbf{u}$ to n items, where $u_i = 1$ if item i is answered correctly, and $u_i = 0$ otherwise. For ability $\boldsymbol{\theta}$, the probability of a correct response is given by

$$P_i(\boldsymbol{\theta}) \equiv P(U_i = 1|\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\mathbf{a}_i'(\boldsymbol{\theta} - b_i\mathbf{1})\right]}\,, \tag{10}$$

[2] If **B** and Γ are unknown, then it is possible to estimate **A** directly, although this form contains additional parameters, and represents a more general model. Estimating parameters of a more general model can in some cases reduce the precision of the estimates. However, this tendancy may be offset by increased sample size.

Table 2: **A**-matrix discrimination value pattern.

	Factor						
	General	Verbal	Math	AR	WK	PC	MK
Item	I	II	III	IV	V	VI	VII
Arithmetic Reasoning							
1	$a_{1,1}$		$a_{3,1}$	$a_{4,1}$			
2	$a_{1,2}$		$a_{3,2}$	$a_{4,2}$			
⋮	⋮		⋮	⋮			
30	$a_{1,30}$		$a_{3,30}$	$a_{4,30}$			
Word Knowledge							
31	$a_{1,31}$	$a_{2,31}$			$a_{5,31}$		
32	$a_{1,32}$	$a_{2,32}$			$a_{5,32}$		
⋮	⋮	⋮			⋮		
65	$a_{1,65}$	$a_{2,65}$			$a_{5,65}$		
Paragraph Comprehension							
66	$a_{1,66}$	$a_{2,66}$				$a_{6,66}$	
67	$a_{1,67}$	$a_{2,67}$				$a_{6,67}$	
⋮	⋮	⋮				⋮	
80	$a_{1,80}$	$a_{2,80}$				$a_{6,80}$	
Math Knowledge							
81	$a_{1,81}$		$a_{3,81}$				$a_{7,81}$
82	$a_{1,82}$		$a_{3,82}$				$a_{7,82}$
⋮	⋮		⋮				⋮
105	$a_{1,105}$		$a_{3,105}$				$a_{7,105}$

where c_i and b_i are the guessing and difficulty parameters, respectively, for item i; $\mathbf{1}$ is a $p \times 1$ vector of 1's; D is the constant 1.7; and $\mathbf{a}_i'$ is a $1 \times p$ vector of discrimination parameters for item i corresponding to the i-th row of $\mathbf{A}$ given by (9). It follows from the assumption of local independence that the probability of a set of observed responses for an examinee of ability $\boldsymbol{\theta}$ is equal to

$$P(U_1 = u_1, U_2 = u_2, ..., U_n = u_n | \boldsymbol{\theta}) = \prod_{i=1}^{n} P_i(\boldsymbol{\theta})^{u_i} Q_i(\boldsymbol{\theta})^{1-u_i} , \tag{11}$$

where $Q_i(\boldsymbol{\theta}) = 1 - P_i(\boldsymbol{\theta})$. The right hand side of (11) is algebraically equivalent to the likelihood function $L(\mathbf{u}|\boldsymbol{\theta})$ where the responses are fixed at observed values.

Next, we shall consider the simultaneous estimation of the full set of p traits denoted by $\boldsymbol{\theta}$—cognizant of the fact that our interests lie in the specification of the first element of $\boldsymbol{\theta}$, namely θ_1 which corresponds to the general ability dimension. Ability estimates for the full vector $\boldsymbol{\theta}$ can be obtained by either maximum likelihood (ML) or Bayesian techniques (Segall, 1996, in press). Given that the population distribution of ability is known, or can be well approximated, Bayesian estimation is often preferable to ML. This is especially

true for short or higher dimensionality tests where item-response data are sparse. Bayesian point estimates of ability can be defined as the mode of the posterior distribution $f(\boldsymbol{\theta}|\mathbf{u})$, which is proportional to the product of the likelihood and prior:

$$f(\boldsymbol{\theta}|\mathbf{u}) \propto L(\mathbf{u}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) .$$

In the current application we shall assume the prior $f(\boldsymbol{\theta})$ to be a multivariate normal density with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$. The modal estimates, denoted by $\hat{\boldsymbol{\theta}}$, are those values of $\boldsymbol{\theta}$ that satisfy the set of p simultaneous equations $\partial \ln f(\boldsymbol{\theta}|\mathbf{u})/\partial\boldsymbol{\theta} = \mathbf{0}$, where

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}|\mathbf{u}) = D \sum_{i=1}^{n} v_i \mathbf{a}_i - \boldsymbol{\theta} ,$$

and

$$v_i = \frac{[P_i(\boldsymbol{\theta}) - c_i]\,[u_i - P_i(\boldsymbol{\theta})]}{(1 - c_i)\,P_i(\boldsymbol{\theta})} .$$

Since there is no closed-form solution to this set of equations, an iterative method must be used, such as the Newton-Raphson method. Suppose we let $\boldsymbol{\theta}^{(m)}$ denote the m-th approximation to the value of $\boldsymbol{\theta}$ that maximizes $\ln f(\boldsymbol{\theta}|\mathbf{u})$, then a better approximation is generally given by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - \boldsymbol{\delta}^{(m)} , \tag{12}$$

where $\boldsymbol{\delta}^{(m)}$ is the $p \times 1$ vector

$$\boldsymbol{\delta}^{(m)} = \left[\mathbf{J}(\boldsymbol{\theta}^{(m)})\right]^{-1} \times \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}^{(m)}|\mathbf{u}) . \tag{13}$$

The matrix $\mathbf{J}(\boldsymbol{\theta})$ is the matrix of second partial derivatives evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$:

$$\mathbf{J}(\boldsymbol{\theta}) = D^2 \sum_{i=1}^{n} \mathbf{a}_i \mathbf{a}_i' w_i - \mathbf{I} ,$$

where

$$w_i = \frac{Q_i(\boldsymbol{\theta})\,[P_i(\boldsymbol{\theta}) - c_i]\,[c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})\,(1 - c_i)^2} .$$

Modal estimates are obtained through successive approximations using (12) and (13) until convergence is reached.

When the joint posterior density $f(\boldsymbol{\theta}|\mathbf{u})$ is multivariate normal, then the marginal distribution of primary interest $f(\theta_1|\mathbf{u}) = \int \cdots \int f(\theta_1, \theta_2, ..., \theta_p|\mathbf{u}) d\theta_2 \cdots d\theta_p$ is itself normally distributed with its expected value equivalent to the corresponding θ_1-mode of the joint posterior distribution (Anderson, 1984, Theorem 2.4.3, p. 31). This observation suggests that for cases in which the posterior is well approximated by a multivariate normal distribution (for moderate to long tests) the first element of $\hat{\boldsymbol{\theta}}$, namely $\hat{\theta}_1$, will provide a point estimate which is approximately equivalent to the expectation of the marginal distribution of θ_1.

Testing Methods

A number of methods for general ability measurement have been applied in practice, or suggested in the literature. Several alternatives are outlined below. These include one traditional approach based on classical test theory, two adaptive testing approaches based on unidimensional IRT, and two new approaches based on multidimensional IRT.

Conventional Number-Right

This procedure is the most commonly used method of measuring general abilities, and is based largely on classical test theory. Typically, test items contained in the battery span several content areas (e.g. math, verbal, spatial, reasoning, etc.). Most commonly, items are scored correct or incorrect, and each examinee receives a total score reflecting the number of correct responses across all scales. The total number-right score is taken as an indicator of general ability. In some instances, subscores are computed for each content area, and the indicator of general ability is taken as a positively weighted sum of the subscores. If the subscales are highly correlated, or if the number of scales is moderately large (10 or more), composites computed from different sets of weights will be highly correlated. In these instances, the choice of weights will tend to have little influence on the precision of the composite score.

Unidimensional Adaptive Testing

Computerized adaptive testing has been demonstrated, in a number of actual and simulated applications (Lord, 1980; Wainer et al., 1990; Weiss, 1978), to provide increased measurement efficiency over conventional paper-and-pencil testing. Although different CAT testing algorithms have been proposed and adopted, most share a common set of core characteristics. Virtually all CAT algorithms are based on unidimensional IRT and employ either the one-, two-, or three-parameter logistic response-model. Items are selected to maximize Fisher or posterior information, and final test scores are obtained using maximum likelihood or Bayesian estimation procedures. In principle, at least two different CAT approaches exist for the measurement of a general test factor.

In one approach, the collection of items spanning the general ability of interest (e.g. math and verbal) are treated as if they form a single pool. These items are calibrated jointly, using for example BILOG (Zimowski, Muraki, Mislevy, & Bock, 1996), and item selection and scoring are based on a unidimensional CAT algorithm. One criticism of this approach is that a fundamental assumption (that of unidimensionality) is violated by the item calibration, selection, and scoring algorithms. If the dimensions spanned by the items from alternative scales are not highly correlated, this violation may reduce CAT efficiency and call into question the important IRT quality of test-score invariance. However, if the dimensions are highly correlated, the information provided by cross-scale responses may improve the efficiency of item-selection and scoring over an approach which ignores this information.

Multi-Unidimensional Adaptive Testing

Another CAT approach to the measurement of general abilities divides the test items into homogenous (unidimensional) item pools, and measures each narrow ability by a sepa-

rate adaptively administered test. For example, in the case of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (Segall & Moreno, 1999), items spanning *general trainability* are divided into four homogenous scales: (a) Arithmetic Reasoning, (b) Word Knowledge, (c) Paragraph Comprehension, and (d) Math Knowledge. Items belonging to each scale were calibrated separately to obtain four pools of unidimensional item parameter estimates. Abilities associated with each scale are measured by a short (15 item) adaptive test. Then, an estimate of general ability is formed from a weighted composite of the four separately estimated ability-values. To the extent that multidimensionality is reduced by forming individual scales, this procedure is preferable to the unidimensional procedure described above. Since however information from cross-scale responses is ignored by the item selection and scoring algorithms, the multi-unidimensional procedure might provide less precise measures of general ability when the dimensions are highly correlated.

Conventional Item Selection with MIRT-Scoring

By applying the multidimensional item response theory (MIRT) scoring algorithms to conventional tests, yet another possibility exists for improving general ability measurement. This approach requires test items be calibrated according to the hierarchical model (8). Then the multidimensional vector of ability estimates can be obtained by the Bayesian procedure (12) and (13), and the estimated value corresponding to the general dimension θ_1 would be taken as a measure of general ability. Lord (1980) illustrates that IRT scoring of conventional unidimensional tests provides higher levels of information than number-right scoring (p. 73). By explicitly modeling the effects of specific test variance on responses, it is likely that an even greater superiority over number-right scoring exists for conventional multidimensional tests scored by MIRT methods. However, it remains to be seen how much improvement in precision can be gained over number-right scoring by using this more computationally intensive MIRT scoring method.

Multidimensional Adaptive Testing

Another MIRT method consists of applying a multidimensional adaptive testing algorithm, such as the one proposed by Segall (1996). This algorithm however may be poorly suited for the current problem since it selects items to maximize precision along all dimensions simultaneously. In the current case, Segall's algorithm would select items to maximize the precision of the nuisance dimensions, as well as the general dimension. Conceivably more precise measurement of the general ability parameter could be achieved by selecting items to maximize its precision directly. van der Linden (in press) has proposed a multidimensional item selection algorithm which minimizes the variance of the ML estimate for a linear combination of abilities. An alternative Bayesian adaptive item selection algorithm is presented below which chooses items to minimize the posterior variance of the general ability parameter.

Suppose that $k-1$ items have already been administered, and the task is to decide which item is to be selected as the next (k-th) item from the set of remaining items R_k. Let the set of administered (or selected) items be denoted by $S_{k-1} = \{i_1, i_2, ..., i_{k-1}\}$, whose elements uniquely identify the items which are indexed in the pool according to $i = 1, 2, ..., I$.

Further, let the set of remaining or candidate items be denoted by the compliment of S_{k-1}, namely $R_k = \{1, 2, ..., I\} \backslash S_{k-1}$.

One approach suggested by Bayesian decision theory is to select the item which minimizes the expected posterior variance of the general ability parameter. An estimate of this variance can be obtained by approximating the posterior with a multivariate normal density based on the curvature at the mode. Specifically, the posterior distribution $f(\boldsymbol{\theta}|\mathbf{u}_k)$ (obtained after the administration of item i and observation of the associated response u_i) can be approximated by a normal distribution having mean equal to the posterior mode $\hat{\boldsymbol{\theta}}^{k-1}$ (calculated from the $k-1$ administered items), and covariance matrix $\Sigma_{i|S_{k-1}}$ equal to the inverse of the posterior information matrix evaluated at the mode $\hat{\boldsymbol{\theta}}^{k-1}$:

$$\Sigma_{i|S_{k-1}} = \left[\Upsilon_{i|S_{k-1}}\right]^{-1} , \tag{14}$$

where the information matrix $\Upsilon_{i|S_{k-1}}$ is minus the expected Hessian (second derivative matrix) of the log posterior

$$\Upsilon_{i|S_{k-1}} = \mathbf{W}_i + \mathbf{I} + \sum_{j \in S_{k-1}} \mathbf{W}_j ,$$

and where $\mathbf{W}_i = D^2 \mathbf{a}_i \mathbf{a}_i' w_i^*$, and

$$w_i^* = \frac{Q_i(\boldsymbol{\theta})}{P_i(\boldsymbol{\theta})} \times \left(\frac{P_i(\boldsymbol{\theta}) - c_i}{1 - c_i} \right)^2 .$$

The posterior information matrix associated with candidate item i is formed from $\mathbf{W}$-terms associated with previously administered items S_{k-1}, and from a $\mathbf{W}$-term associated with the candidate item i. Given that the posterior distribution is normal or nearly so, the first (upper-left) diagonal element of $\Sigma_{i|S_{k-1}}$ will provide a suitable approximation to the variance of the marginal posterior distribution of θ_1.

To implement this approach, MIRT item parameters must be specified according to the parameterization given by (8). To select the k-th item, the posterior covariance matrix (14) is calculated for each candidate item—the item associated with the smallest variance-term in the first (upper-left) diagonal element is selected for administration. The final ability estimate for the general dimension is taken as the $\hat{\theta}_1$-element from the joint posterior mode $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_p\}'$ estimated from all responses according to (12) and (13).

Simulation Study

Compared to conventional testing methods, there are compelling theoretical benefits for the application of MIRT scoring and item selection procedures to general ability measurement. However, it remains to be seen if some of the model assumptions (i.e. asymptotic normality of the posterior distribution) hold sufficiently well to produce the intended results. It also remains to be seen if the benefits of the MIRT procedures (increased precision and reduced test-lengths) outweigh the additional computational complexities. A comparison of alternative testing methods is made using simulated data based on a large scale high-stakes selection battery: The Armed Services Vocational Aptitude Battery (ASVAB). This battery, which has been shown to possess a large general-ability component (Ree & Earles, 1994) is used to qualify applicants for military service and training programs.

Table 3: Item source for simulation pool.

Subtest	No. of Items per Form	No. of Forms	Total
AR: Arithmetic Reasoning	30	4	120
WK: Word Knowledge	35	4	140
PC: Paragraph Comprehension	15	4	60
MK: Math Knowledge	25	4	100
Total	105		420

Item Pool

An artificial item pool of 420 items was modeled from the combination of four paper-and-pencil forms of the ASVAB. The composition of the pool is displayed in Table 3. The tri-factor pattern of discrimination parameters assumed for each of the four ASVAB forms is specified in Table 2. Consistent with the hierarchical model (9), each item has three nonzero discrimination values, one for the general factor, another for the math or verbal factor, and a third for a subtest specific factor. For example, Arithmetic Reasoning items load on the General factor (I), the Math factor (III), and an AR specific factor (IV). Thus each AR item will possess at most three nonzero discrimination parameters among the seven possible. A similar interpretation can be made for the other item types (WK, PC, and MK). This pattern of constraints was assumed for each of the four ASVAB forms.

The IFACT item-parameter estimation procedure (Segall, 1998) was used to specify item response functions (IRFs) according to the assumed tri-factor pattern. This procedure performs a confirmatory IRT item-factor analysis, where the pattern of free and fixed item loadings (discrimination values) is specified *a priori*. The IFACT procedure is based on an extension of the Markov Chain Monte Carlo method proposed by Albert (1992). The procedure expands Albert's approach from a single latent dimension to multiple latent dimensions, allows user-specified discrimination parameters to be constrained to zero, and incorporates a provision for estimating a guessing parameter. The procedure assumes that the distribution of latent abilities is multivariate normal with mean vector $\mathbf{0}$ and unit-diagonal covariance matrix $\mathbf{I}$.

Two datasets (live and simulated) were used to mimic the situation where item-responses are modeled by one set of true IRFs, but scoring and item selection are based on another, possibly misspecified set of IRFs. The datasets and associated parameter estimates are described below.

Live calibration data. Live calibration data for each of the four forms were gathered from 12,000 applicants taking the ASVAB to qualify for service in the U. S. Military. Each test-record contained responses to 105 items of either Form 1, Form 2, Form 3, or Form 4. The numbers of items from each content area (subtest) was fixed across forms—each form containing 30 AR items, 35 WK items, 15 PC items, and 25 MK items. Four separate multidimensional IFACT calibrations were conducted, one for each form. Each calibration placed constraints on the pattern of free and fixed discrimination parameters summarized in Table 2. These parameters, denoted by IFACT(U), were treated as true (population) values for generating item responses. Table 4 provides the means and standard deviations

Table 4: Descriptive statistics for live data item parameters.

	Discrimination Parameters a_i								
Statistic	General I	Verbal II	Math III	AR IV	WK V	PC VI	MK VII	Difficulty b	Guessing c
Arithmetic Reasoning									
Mean	1.19		0.33	0.21				0.04	0.22
SD	0.48		0.39	0.34				0.68	0.08
Word Knowledge									
Mean	0.91	0.67			0.17			−0.60	0.22
SD	0.29	0.42			0.37			0.93	0.06
Paragraph Comprehension									
Mean	0.74	0.37				0.32		−0.52	0.21
SD	0.30	0.23				0.22		0.59	0.05
Math Knowledge									
Mean	1.03		0.59				0.46	0.09	0.21
SD	0.32		0.46				0.56	0.51	0.08

(SD) of discrimination, difficulty, and guessing parameters for each content area.

Simulated calibration data. Four response sets were simulated from the corresponding four-sets of multidimensional item parameter estimates obtained from live data. These datasets were of the same form as the live-response datasets (four randomly equivalent groups of 3000 respondents, 105 items per test, and ability $\boldsymbol{\theta}$ sampled from $N(\mathbf{0}, \mathbf{I})$. Three different sets of item-selection/scoring parameters were obtained by applying different estimation approaches to the simulated calibration data:

1. BILOG (Zimowski et al., 1996) was used to estimate unidimensional item parameters from the collection of all 105 items of each form. These four sets of parameter estimates were combined to form a single pool of 420 items.
2. BILOG was also used to estimate unidimensional 3PL item parameters for each content area of each form, which resulted in 16 (= 4 forms × 4 content areas) separate sets of parameter estimates. These 16 sets of parameter estimates were combined into four pools of items—one pool for each content area: AR, WK, PC, and MK. Pool sizes are listed in the last column of Table 3.
3. IFACT was used to estimate multidimensional item parameters using the same design and discrimination-parameter constraints applied to the live datasets. The resulting four sets of parameter estimates were combined to form a single pool of 420 items. Root mean squared differences ([RMSD]; Hulin, Lissak, & Drasgow, 1982) were computed between this set of IRFs (obtained from simulated calibration data) and the true set (obtained from the live calibration data), where the pairs of IRFs defined by (10) were evaluated at the 3,000 true ability values used to generate the simulated calibration data. The average RMSD across all 420 items was 0.027.

Conditions

To evaluate the precision of the alternative item selection and scoring strategies, conditional distributions of test-scores for fixed levels of the general dimension were obtained. For each of the five testing methods described below, 500 replications were conducted at each of 31 equally spaced ability levels $-3, -2.8, -2.6, ..., +2.8, +3$ along the general dimension (θ_1). For each replication, ability values for the secondary nuisance dimensions $(\theta_2, \theta_3, ..., \theta_p)$ were sampled from a multivariate normal distribution with mean vector $\mathbf{0}$, and unit-diagonal covariance matrix. The 500 resulting test scores x for each level of θ_1 were used to estimate the first two moments of the conditional score distributions. These moments are denoted by $\mathrm{m}(x|\theta_1)$ and $\mathrm{Var}(x|\theta_1)$ for the conditional mean and variance, respectively.

The conditional test-score moments were examined for five different testing methods. In each evaluation described below, the multidimensional item-parameters obtained from live data IFACT(U) were used to generate responses. Parameters used for item selection and scoring (if required) were obtained from values estimated by Approach 1, 2, or 3. The adaptive testing methods consisted of fixed length tests (totaling 60-items) which was about half the number of items administered by the conventional nonadaptive testing methods. Based on previous research with unidimensional adaptive testing methods (e.g. Green, 1983; McBride & Martin, 1983; Sands, Waters, & McBride, 1997; Urry, 1977) we would expect the shorter adaptive tests to achieve about the same (or slightly higher) level of precision as the conventional testing methods. Additional details of the simulation for each testing method are provided below.

Conventional number-right (CONV-NR). This condition modeled number-right scoring applied to a conventional administration of Form 1 of the ASVAB. For each simulated test-taker, dichotomous responses were generated by evaluating the item response function (10) at the true ability level and comparing the probability value to a pseudo random uniform number. A total number-right score was calculated from the sum of the 105 dichotomous item-scores.

Unidimensional adaptive testing (CAT-UNI). This condition modeled the application of unidimensional adaptive testing algorithms to the multidimensional item pool. The item pool consisted of all 420 items whose observed parameters were specified by Approach 1. Item selection was based on ML information (Lord, 1980, Section 10.2), where each candidate item is evaluated at the provisional ability estimate. Provisional and final scoring was based on unidimensional Bayesian posterior modal estimates, assuming a standard normal prior. Fixed-length tests of 60 items were simulated for each test-taker. The posterior mode based on all adaptively administered items was taken as an estimate of general ability.

Multi-unidimensional adaptive testing (CAT-MUNI). This condition modeled the application of unidimensional adaptive testing algorithms to unidimensional—or nearly unidimensional item pools. Four separately administered adaptive tests of 15 items each were simulated for each test-taker, one test for each content area: AR, WK, PC, and MK. The four pools of item parameters used for item-selection and scoring were specified by Approach 2. Item selection was based on ML information, and scoring was based on Bayesian posterior modal estimates. A final score, taken to be an estimate of general

ability, was formed from the unit-weighed sum of the four modal estimates for each content area: $\hat{\theta}_{\mathrm{AR}} + \hat{\theta}_{\mathrm{WK}} + \hat{\theta}_{\mathrm{PC}} + \hat{\theta}_{\mathrm{MK}}$.

Conventional item selection with MIRT-scoring (CONV-MIRT). This condition modeled MIRT scoring applied to a conventional administration of Form 1. For each simulated test-taker, dichotomous responses were generated for the 105 items comprising Form 1. Seven-dimensional Bayesian modal estimates were calculated from the parameters of Approach 3 corresponding to Form 1 items. The modal estimate $\hat{\theta}_1$ corresponding to the general dimension was taken as an estimate of general ability.

Multidimensional adaptive testing (GMAT). This condition modeled the application of the multidimensional adaptive testing algorithm given by (14), where items are selected to minimize the posterior variance of the general ability parameter θ_1. Parameters for item selection and scoring were specified by Approach 3. Fixed-length tests of 60 items were simulated for each test-taker. The modal estimate $\hat{\theta}_1$ corresponding to the general dimension of the joint posterior mode (based on all 60 responses) was taken as an estimate of general ability.

Results

For each testing method, two different precision summaries were calculated: (a) a score information function, and (b) a reliability index. An additional analysis examined the distribution of content across ability ranges for the GMAT (multidimensional adaptive testing) approach.

Score information. For each testing method, the information function of the associated scoring formula x (Birnbaum, 1968, Section 17.7)

$$I(\theta_1, x) = \frac{\left[\frac{d}{d\theta_1}\,\mathrm{m}(x|\theta_1)\right]^2}{\mathrm{Var}(x|\theta_1)} \tag{15}$$

was approximated from the first two moments of the conditional score-distributions. The numerator of (15) was approximated by a cubic smoothing spline fit to the 31 conditional means $\mathrm{m}(x|\theta_1 = -3), \mathrm{m}(x|\theta_1 = -2.8), ..., \mathrm{m}(x|\theta_1 = 3)$ using an algorithm described by de Boor (1978, p. 235–243). Since the cubic spline approximation is a piecewise polynomial of order 4, the required derivatives were easily obtained by evaluation of the derivative of the appropriate piecewise polynomial. The denominator of (15) was also approximated by a smoothing spline fitted to the conditional variances. The resulting score information functions are displayed in Figure 1. As indicted, the lowest level of information is observed for the conventional number-right (CONV-NR) testing method. MIRT scoring of conventional tests (CONV-MIRT) approximately doubles the level of information obtained from conventional item responses. Both methods of unidimensional adaptive testing (CAT-UNI and CAT-MUNI) demonstrate gains in precision over conventional number-right testing, but fall somewhat sort of the level achieved by CONV-MIRT over the middle and lower ability ranges. By far, the highest levels of information were observed for the multidimensional adaptive testing procedure (GMAT) which displayed a several-fold increase in information over competing methods.

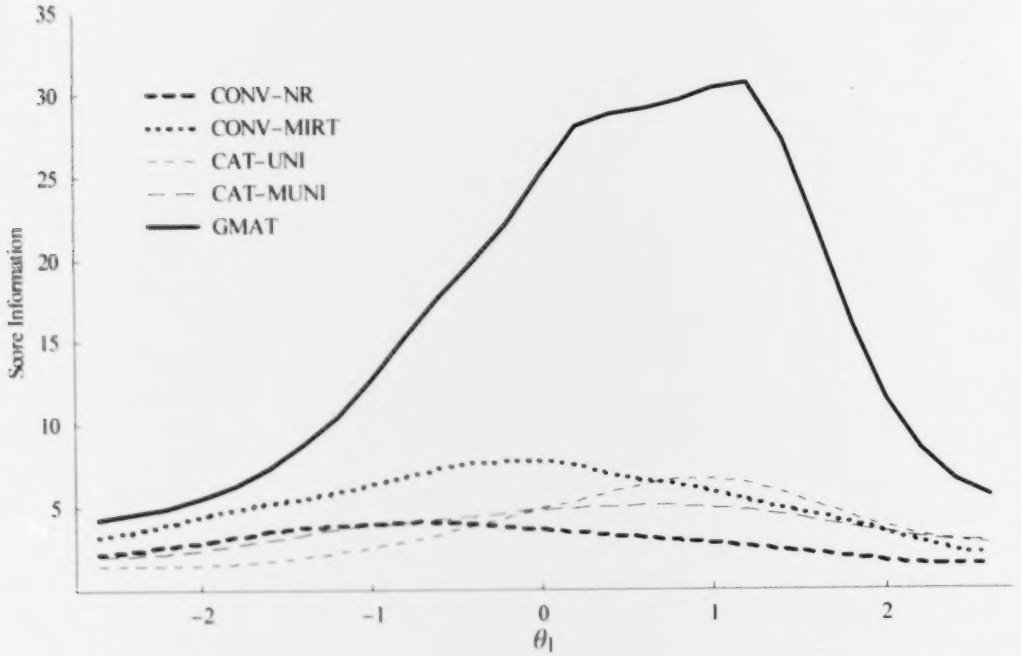

Figure 1. Score Information Functions.

Reliability. The second precision measure η^2 provides an estimate of the proportion of variance of the test score x that can be predicted from θ_1:

$$\eta^2 = \frac{\text{Var}\left[\text{m}\left(x|\theta_1\right)\right]}{\text{Var}\left[\text{m}\left(x|\theta_1\right)\right] + \text{E}\left[\text{Var}\left(x|\theta_1\right)\right]} ,$$

where

$$\text{Var}\left[\text{m}\left(x|\theta_1\right)\right] = \sum_{k=1}^{31} w_k \left[\text{m}\left(x|\theta_1 = L_k\right) - \text{E}\left(x\right)\right]^2 ,$$

$$\text{E}\left(x\right) = \sum_{k=1}^{31} w_k \, \text{m}\left(x|\theta_1 = L_k\right) ,$$

$$\text{E}\left[\text{Var}\left(x|\theta_1\right)\right] = \sum_{k=1}^{31} w_k \, \text{Var}\left(x|\theta_1 = L_k\right) ,$$

and where the w_k's are proportional to the height of the normal density evaluated at $L_1 = -3$, $L_2 = -2.8$, ..., $L_{31} = 3$. The w_k's are normalized so $\sum_k w_k = 1$. This index provides

Table 5: Reliability indices.

	Condition				
	CONV-NR	CAT-UNI	CAT-MUNI	CONV-MIRT	GMAT
Test Length	105	60	60	105	60
η^2	.77	.80	.81	.86	.95

Table 6: Item content distribution by ability level (GMAT).

	Ability Range				
Content	$-3 < \theta < -1$	$-.8 < \theta < -.4$	$-.2 < \theta < .2$	$.4 < \theta < .8$	$1 < \theta < 3$
AR	36.5	49.2	54.5	56.2	54.2
WK	26.6	13.6	10.6	9.7	12.4
PC	15.0	14.5	12.0	10.4	11.6
MK	21.9	22.7	22.9	23.7	21.8
Total	100.0	100.0	100.0	100.0	100.0

an estimate of the reliability of observed test scores for a population with $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I})$. The reliability values and test-lengths for each of the five conditions are displayed in Table 5. As indicated, the conventional administration method scored by number-right (CONV-NR) displays the lowest reliability. Both forms of unidimensional adaptive testing (CAT-UNI and CAT-MUNI) display moderately higher values, with conventional administration scored by MIRT (CONV-MIRT) displaying large gains in reliability over the other three methods. The highest level of reliability is achieved by the multidimensional adaptive testing method (GMAT) which displays near perfect measurement of the general dimension ($\eta^2 = .95$).

Content usage. Table 6 displays the distribution of content across five ability ranges for the multidimensional adaptive testing approach (GMAT). As indicated, the balance of math and verbal items appears to shift between the lower and higher ability ranges. Over the lowest range, verbal items account for about 40% of the administered items. Over the highest ability range, the percentage of administered verbal items falls to about 24%. This is consistent with the mean difficulty value displayed for WK in Table 4, which indicates that a large portion of the WK items discriminate over the lower ability range. On average, AR accounts for about half of the administered items, MK for about 23-percent, and WK and PC for about 13-percent each. The predominance of AR items is consistent with their high average discrimination parameter value for the general dimension, and somewhat lower loading values on the nuisance dimensions, as displayed in Table 4.

Conclusions and Discussion

The results presented here indicate that substantial gains in the measurement efficiency of a general test factor can be achieved by application of the proposed MIRT strategies. These strategies parameterize the hierarchical ability model in terms of uncorrelated general and specific abilities. Bayesian MIRT estimation of the general ability parame-

ter for a conventional test (using the appropriate model) displays a significant increase in precision over number-right scoring methods, and a moderate increase in precision over unidimensional adaptive testing methods. Substantial gains in measurement efficiency are observed for the multidimensional adaptive testing approach which selects items to minimize the posterior variance of the general ability parameter. Using this approach, a several fold increase in information can be achieved over unidimensional adaptive and conventional testing strategies, resulting in near perfect measurement of the general test factor.

One factor precipitating the dramatic increase in precision by MIRT methods is the poor performance of conventional and unidimensional IRT methods for measuring the general test dimension. For example, conventional administration with number-right scoring (CONV-NR) produced a reliability estimate of just 0.77. Note that this estimate is substantially lower than published measures of ASVAB reliability (of about 0.92) based on a similar population of examinees (Palmer, Hartke, Ree, Welsh, & Valentine, 1988). This discrepancy is likely due to the way in which the nuisance variance influences precision estimates such as alternate forms reliability coefficients. Because each nuisance dimension spans across forms, it is likely that a large portion of the resulting covariance between forms is inappropriately classified as true-score variance, thus inflating the reliability estimate. A similar argument can be made for the discrepancy between observed and published information functions for conventional and unidimensional CAT testing methods. Unless information is examined in the context of the appropriate MIRT model, covariation in item performance caused by the nuisance dimensions will likely lead to inappropriately inflated precision estimates.

One interesting finding arrises from an examination of the mixture of item content administered as a function of general ability level (Table 6). In general, we would expect the usage rates of item-types constituting a large portion of the item pool (last column of Table 3) to be higher than those with a smaller representation in the pool. The rationale being that among larger item-groups, there are likely to be more items of high information or appropriate difficulty than among a smaller item-group. Compared to their proportional representation in the pool, AR items appear to be over-administered by a ratio of about 2 : 1, while the number of WK items appears to be under represented by a ratio of about 1 : 2. The numbers of administered PC and MK items appear proportional to their representation in the pool. The other content-related finding of interest is the shift in administration frequency of AR and WK items across ability levels: WK items were administered more frequently to low ability test-takers, while AR items were administered more frequently to moderate and high ability test-takers. These findings suggest AR items are a more useful measure of general ability than the other item-types studied here, while the relative usefulness of WK and AR items depends on ability level. Additional studies of other pools would be required to determine if this finding is a general trend related to item-content, or is unique to the specific ASVAB item-pool examined.

Before the proposed MIRT item selection and scoring methods are routinely applied, several areas of investigation would be productive. First, the development of a confirmatory item-factor analytic procedure for testing the fit of alternative hierarchical battery structures would be useful—one which used all the information contained in the pattern of responses. Existing confirmatory item-factor analytic methods model item-covariances, and thus provide less information than might otherwise be achieved. Second, multidimensional item calibration using such methods as the IFACT procedure should be further investigated

to examine the sample sizes required for sufficiently accurate item-parameter estimation. Smaller sample-size requirements would facilitate the application of the proposed MIRT approach. Third, the usefulness of conventional non-adaptive tests for measuring general ability might be significantly enhanced by the expansion of existing test-assembly procedures (e.g. Swanson & Stocking, 1993; van der Linden & Boekkooi-Timminga, 1989) to incorporate target precision-levels of the general test-factor into the objective function for choosing items. Typical test-assembly models applied to general aptitude batteries choose items to maximize precision along narrowly defined sub-dimensions without consideration of the items' effect on the measurement of a general test factor.

The proposed MIRT methods can be applied to less general areas of knowledge or ability than those applications involving "general mental ability." For example, when measuring foreign language proficiency, reading, writing, speaking, and listening may be treated as nuisance dimensions with a general factor underlying performance in these four areas. This underscores the fact that caution should be exercised in interpreting the general factor measured—this factor may not necessary be equivalent to the general cognitive ability of interest since its nature depends on the tests entered into the battery (Jensen, 1998, Chapter 4). For example, the general factor estimated from a battery consisting of only verbal tests would contain specific (verbal) variance, as well as *g*. The general ability estimate obtained by the proposed MIRT methods is dependent on both: (a) the general factor that spans all tests contained in the battery, and (b) the specific sources of test-variance that in effect define and eliminate unwanted sources of variance.

The results presented here indicate that substantial gains in precision for the measurement of a general test factor can be achieved by application of the proposed MIRT item-selection and scoring algorithms. If this result holds for other test-batteries, the prediction of learning and performance can be significantly enhanced in a wide variety of instances in which standardized tests are used.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* New York: John Wiley & Sons.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). Evanston, NY: Harper & Row.

de Boor, C. (1978). *A practical guide to splines.* New York: Springer-Verlag.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 69–80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: The University of Chicago Press.

Hattie, J. (1981). *Decision criteria for determining unidimensionality.* Unpublished doctoral dissertation, University of Toronto, Canada.

Holzinger, K. J. (1937). The bi-factor method. *Psychometrika, 2,* 41–54.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont, CA: Wadsworth Publishing Company.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249–260.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning.* New York: Plenum.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications.* New York: Wiley.

Humphreys, L. G. (1986). Describing the elephant. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* Norwood, NJ: Ablex.

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75,* 28–42.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Palmer, P., Hartke, D. D., Ree, M. J., Welsh, J. R., & Valentine, L. D. (1988). *Armed Services Vocational Aptitude Battery (ASVAB): Alternative forms reliability (Forms 8, 9, 10 and 11)* (AFHRL-TP No. 87-48). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Ree, M. J., & Earles, J. A. (1991a). Predicting training success: Not much more than g. *Personnel Psychology, 44,* 321–332.

Ree, M. J., & Earles, J. A. (1991b). The stability of g across different methods of estimation. *Intelligence, 15,* 271–278.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current directions in psychological science, 1,* 86–89.

Ree, M. J., & Earles, J. A. (1994). Predicting job success: Not much more than g. *Journal of Applied Psychology, 79,* 518–524.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Segall, D. O. (1998). IFACT computer program Version 1.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation [Computer program]. Seaside, CA: Defense Manpower Data Center.

Segall, D. O. (in press). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Boston: Kluwer-Nijhoff.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151–166.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: The University of Chicago Press.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181–196.

van der Linden, W. J. (in press). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54*, 237–247.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (Ed.). (1978). *Proceedings of the 1977 computerized adaptive testing conference*. Minneapolis, MN: University of Minnesota.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3*, 23–40.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintence for binary items*. Chicago: Scientific Software.